

Scientific Big Data Analytics

Practice & Experience



Dr - Ing. Morris Riedel et al.

Adjunct Associated Professor, University of Iceland, Iceland

Juelich Supercomputing Centre, Germany

Head of Research Group High Productivity Data Processing



The International Conference on Cloud and Autonomic Computing (CAC 2014)

Imperial College, London September 8-12, 2014



Research Field Key Technologies

Jülich Supercomputing Centre

Supercomputing & Big Data



UNIVERSITY OF ICELAND

SCHOOL OF ENGINEERING AND NATURAL SCIENCES

**FACULTY OF INDUSTRIAL ENGINEERING,
MECHANICAL ENGINEERING AND COMPUTER SCIENCE**



'Big Data' in Science & Engineering
Scientific Big Data Analytics
Selected Applications & Experiences
Key Examples of Analytics Practices
Analytics Tools – Lessons Learned
The Role of International Activities
Some Questions & Possible Answers



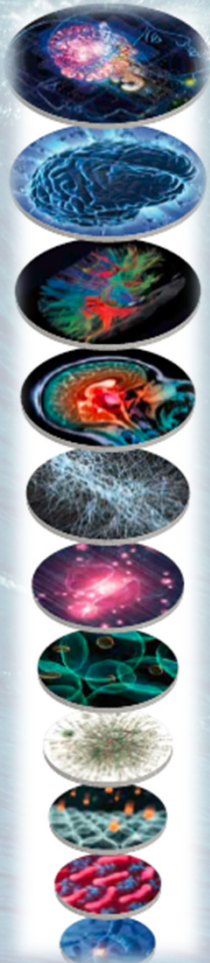
'Big Data Waves'

Volume

Variety

Velocity

Context



Understanding 'Big Data Waves'

Big Data Streams with 'high velocity' ...



'Web Actions'

Data stream with **data** 'hidden' in logs/events

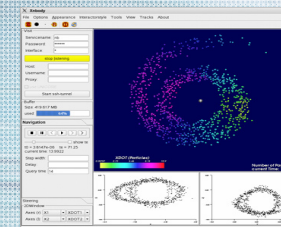
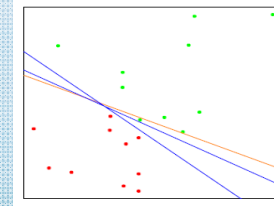
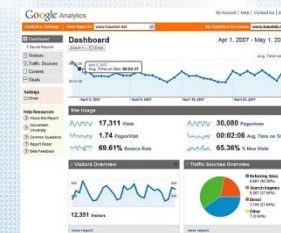
*n optional
filters*

'Measurement Devices'

Data stream with measured **data**

'Computational Simulations'

Data stream with **data** of HPC/HTC simulation

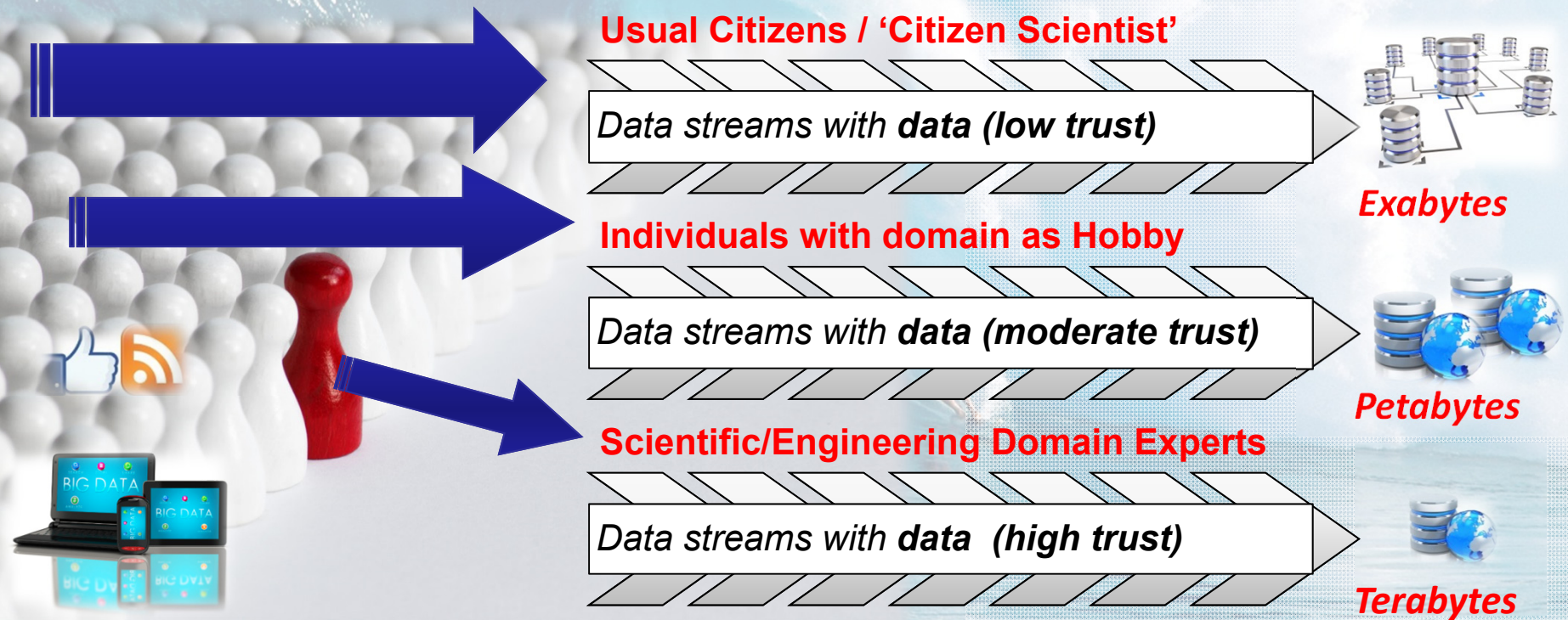


← Data Sources →

← Data Sinks →

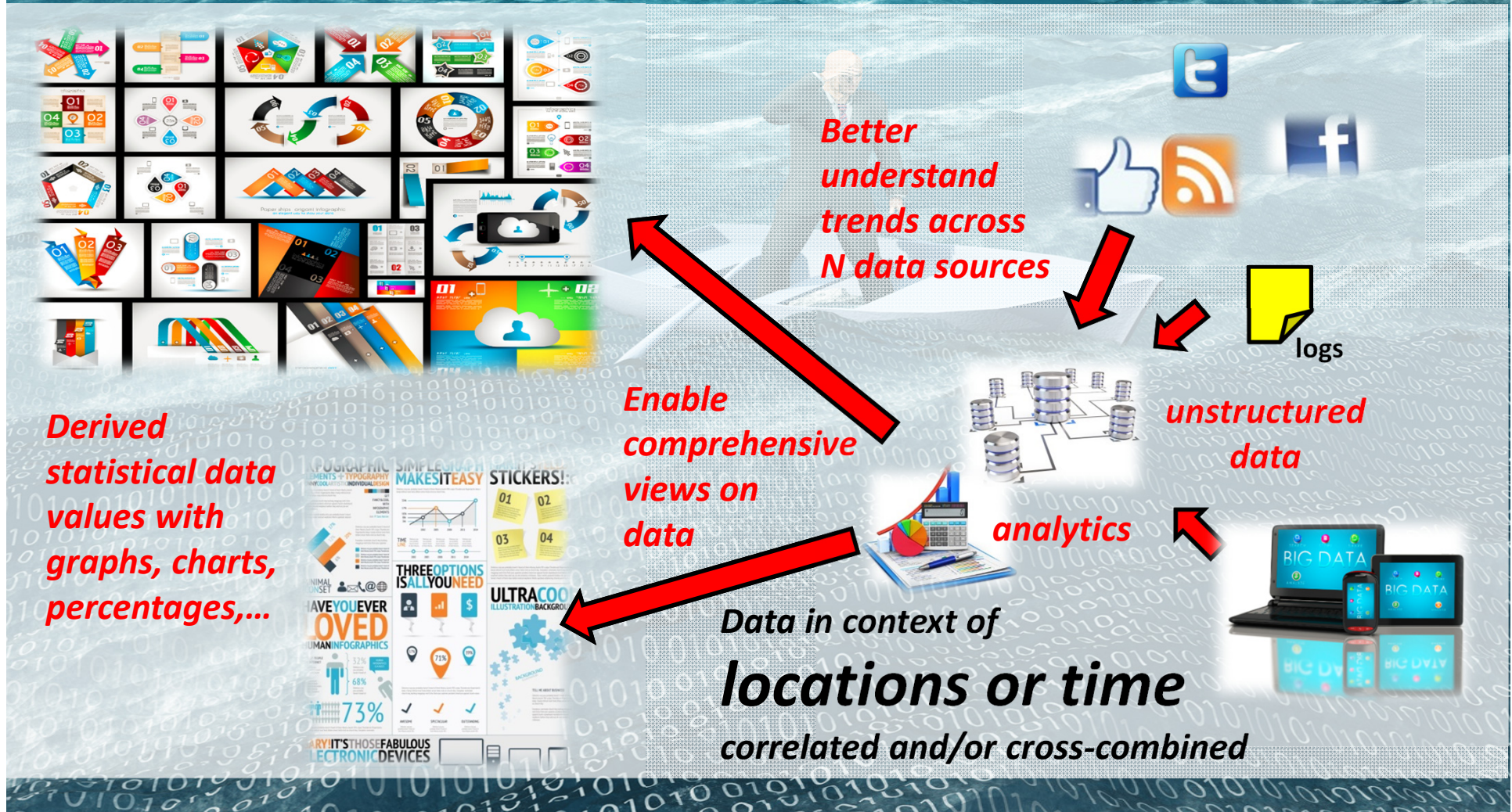
'Crowdsourcing'...

...increases # of Big Data Streams



Infographics

Compact Combination of many Data Visualizations





- ❖ ...
- ❖ **Online Social Media**
(videos, blogs, tweets,...)
- ❖ **Large number of log files**
(Web server log, call center log,...)
- ❖ **Communication data**
(E-Mails, chats, notes, letters, ...)
- ❖ **Various document formats**
(spreadsheet, presentation, docs)
- ❖

Most data in the world...

... is 'unstructured'



**Text
Analytics**

**Data
Mining**

Keep for 'future unknown use'

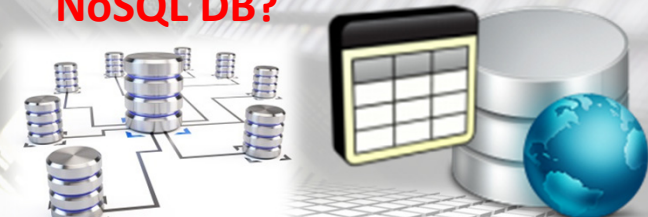
NoSQL DB?

SQL DB?

In-memory?

Disks?

Tapes?



New Forms of Data Structures with NoSQL

Optimized for 'write/once' & 'read/many' or 'In-Memory'



Selected Features

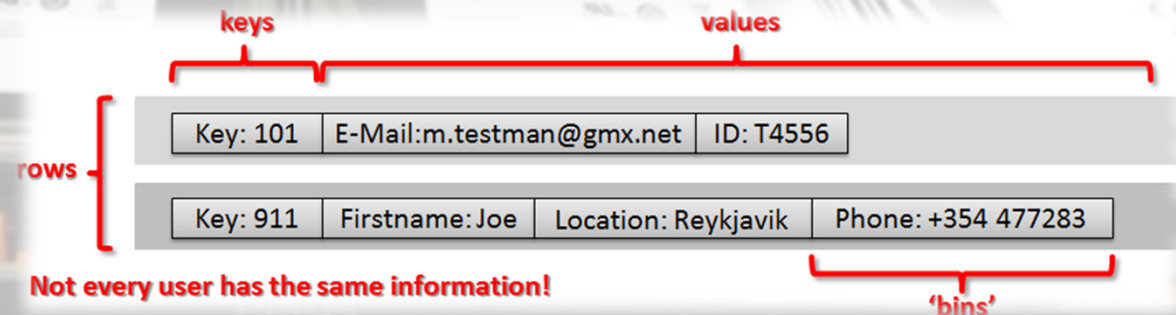
- Simplicity of design and deployment
- Horizontal scaling
- Less constrained consistency models
- Finer control over availability
- Simple retrieval and appending

...

Types

- Key-Value-based (e.g. Cassandra)
- Column-based (e.g. Apache Hbase)
- Document-based (e.g. MongoDB)
- Graph-based (e.g. Neo4J)

'String-based Key-Value Stores' used today



Big Data Waves – Surfboards – Breakwaters

How to engage in the rising tide of ‘Scientific Big Data’?

Unsolved Questions:

Scale
Heterogeneity
Stewardship
Curation
Long-Term Access and Storage

Research Challenges:

Collection, Trust, Usability
Interoperability, Diversity
Security, **Smart Analytics**,
Education and training
Data publication and access
Commercial exploitation
New social paradigms
Preservation and sustainability



[1] *Riding the Wave*,
EC Report, 2010



[2] *A Surfboard for riding the wave*,
Report, 2012

**‘Time for Concrete’
Next Steps →**



Smart Analytics are Needed to Take Advantage of Big Data

The challenge is to understand which analytics make sense

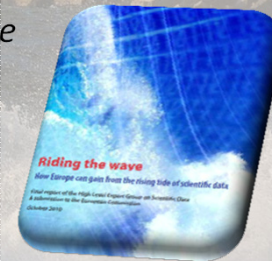
'Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all cross-disciplinary problems that require high-performance data storage, **smart analytics**, transmission and mining to solve.'



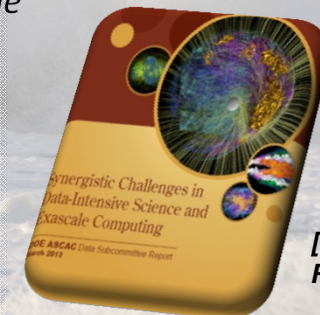
'In the data-intensive scientific world, **new skills are needed for** creating, handling, **manipulating, analysing**, and making available large amounts of data for re-use by others.'

[2] A Surfboard for riding the wave, Report, 2012

'**Integration of data analytics** with exascale simulations represents a new kind of workflow that will impact both **data-intensive science and exascale computing**.'



[1] Riding the Wave, EC Report, 2010



[3] DoE ASCAC Report, 2013

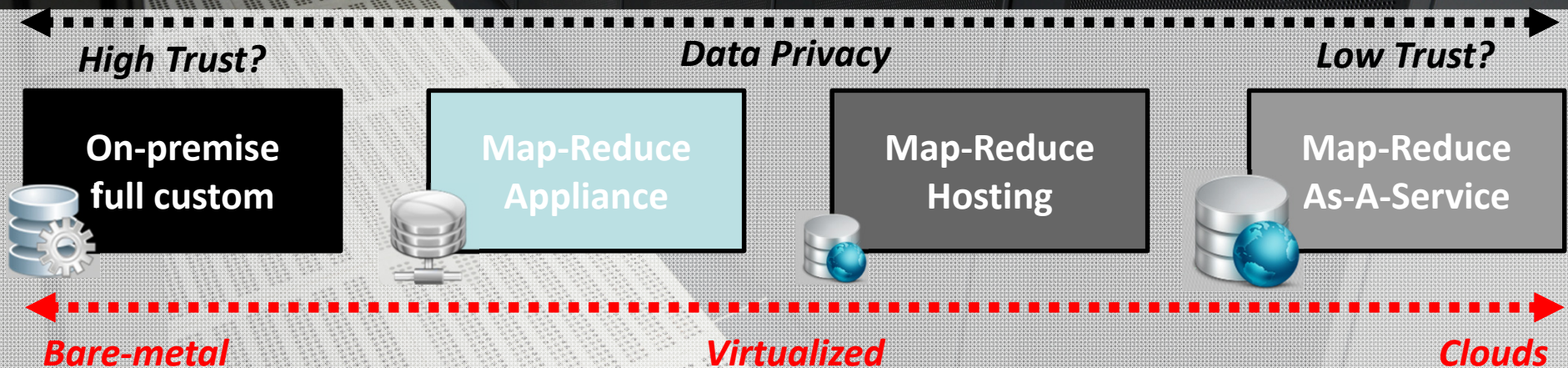


Smart options to move 'data to strong computing power' ...

... or move 'compute tasks close to data'

Total Cost of Ownership vs. ...

... Pay per Use

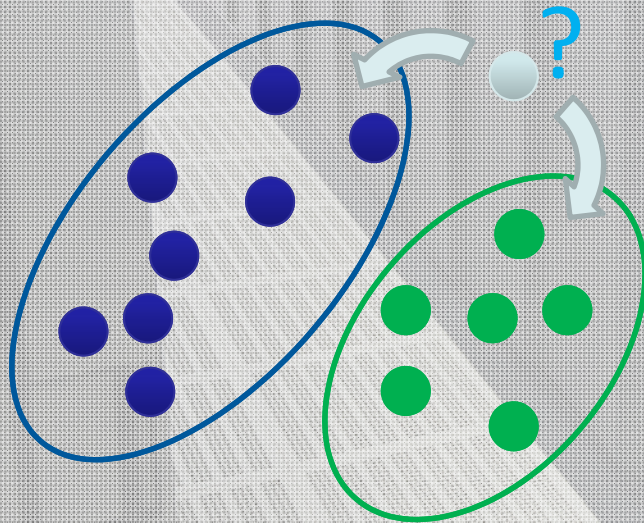


[9] Inspired by a study on Hadoop by Accenture

Making use of Big Data

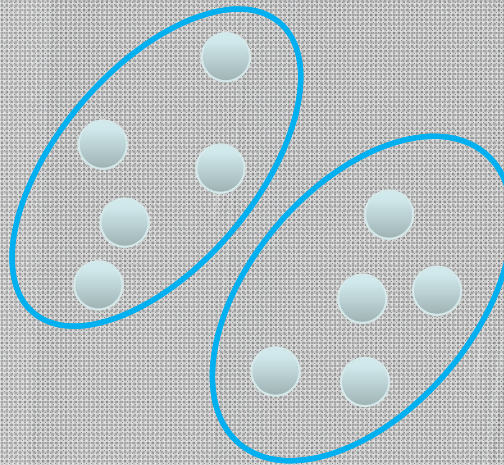
Applying 'smart data analytics' techniques

Classification



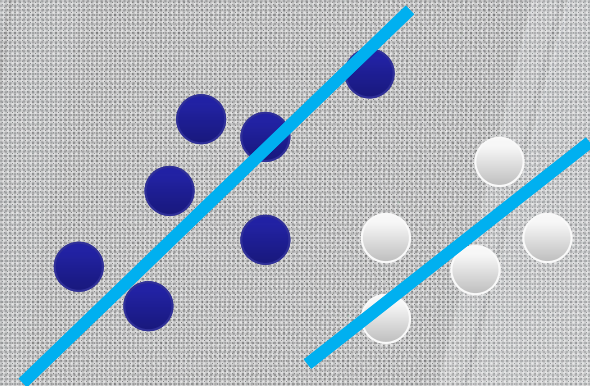
- ✓ Groups of data exist
- ✓ New data classified to existing groups

Clustering



- ✓ No groups of data exist
- ✓ Create groups from data close to each other

Regression

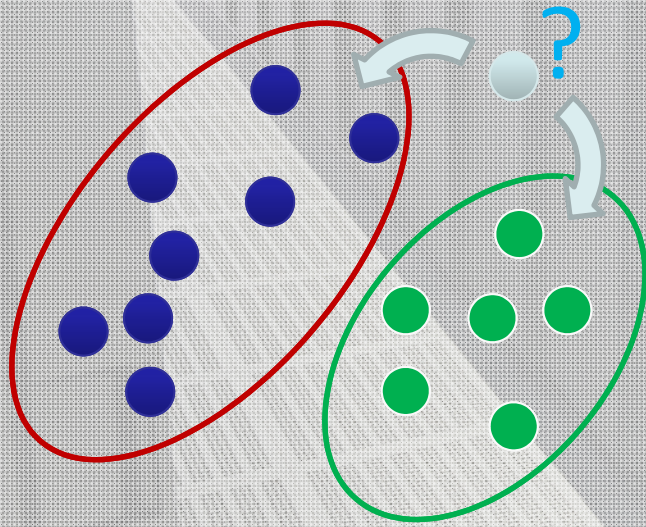


- ✓ Identify a line with a certain slope describing the data

Predictive Analytics Example

Apply 'collaborative filtering' techniques

Classification



Recommender Systems Increase Revenue

Movies		
ID	Name	Category
101	The Internship	Comedy
102	The Purge	Thriller
103	Much About Nothing	Comedy
104	Dirty Wars	Drama
105	Wish You Were Here	Drama
106	Syrup	Comedy
107	Fast & Furious	Thriller

Customers	
ID	Name
1	Mike Jones
2	Steve Thomas
3	Peter Sloan
4	Vish Macnhi
5	Tim Albright

Transactions		
Customer ID	Movie ID	Rating
1	101	
1	102	3
1	103	2.5
2	101	2
2	102	2.5
2	103	
2	104	2
3	101	2.5
3	104	4
3	105	4.5
3	107	
4	101	
4	103	3
4	104	4.5
4	106	4
5	101	4
5	102	3
5	103	2
5	104	4
5	105	3.5
5	106	4

LEGEND	
	1
	2
	3
	4
	5



Scalable & Parallel

Item	User 1	User 2	User 3	User 4	User 5
101	4	2	2.5	5	4
102	3	2.5			3
103	2.5			3	2
104		2	4.5	4.5	4
105			4.5		3.5
106				4	4
107			3		

Recommendation

[16] Apache Mahout Tutorial, YouTube Video



Using
Open Source
Tools



Past History Space

X

Recommendation
System

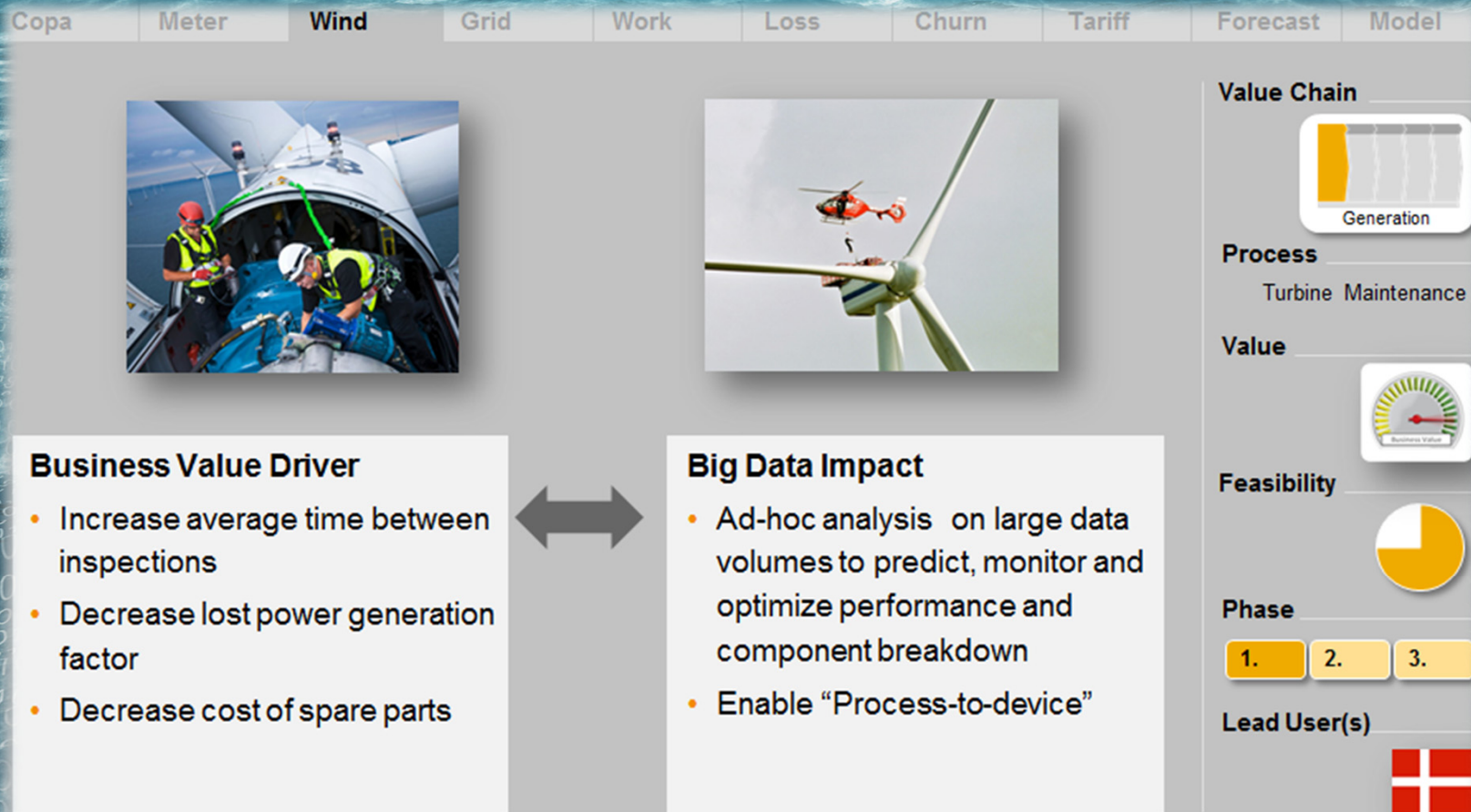
Different Techniques

Prediction Space

y

Utilities Sector Industry Reference

Instant Maintenance Workforce Management



➤ Slide courtesy of Dr. S. Fischer, Global Head of Applied Research – SAP AG (now working at Triumph)

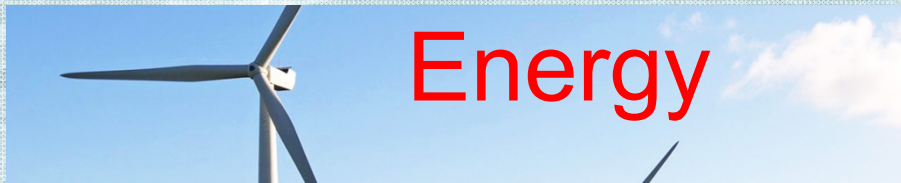
Smart Data Innovation Lab

Companies and Academia work in Focussed Areas

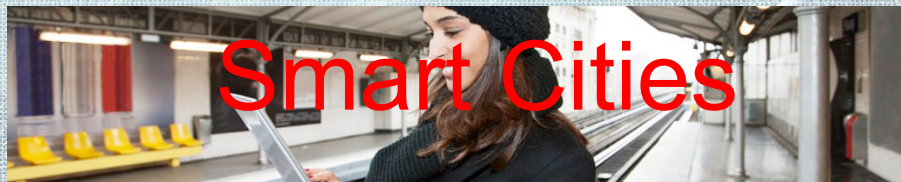
Industry 4.0



Energy



Smart Cities



Personalised Medicine



ABB

ABB AG Forschungszentrum

BASF

BASF SE

BAYER

Bayer Technology Services GmbH

Carl von Ossietzky Universität Oldenburg

Carl von Ossietzky Universität Oldenburg, Abt. Wirtschaftsinformatik I / VLBA

DFK

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

dkfz

Deutsches Krebsforschungszentrum

blueyonder

BOGEN

BOSCH

Carl von Ossietzky Universität Oldenburg

BOGEN Electronics GmbH

Robert Bosch GmbH

Fraunhofer IAIS

Fraunhofer IAO

JÜLICH FORSCHUNGSZENTRUM

CAS

CAS Software AG

EBID

EBID Service AG

echobot

echobot Media Technologies GmbH

Fraunhofer Institut Intelligente Analyse- und Informationssysteme IAIS

Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

Forschungszentrum Jülich GmbH

econdan

econdan GmbH

EnBW

EnBW Energie Baden-Württemberg AG

HITACHI

Hitachi Data Systems GmbH

FZI

Forschungszentrum Informatik

KIT

Karlsruher Institut für Technologie (KIT)

RWTH AACHEN UNIVERSITY

RWTH Aachen Joint Research Center for Computational Biomedicine

IBM

IBM Deutschland GmbH

Infineon

Infineon Technologies AG

leserauskunft

Leserauskunft GmbH

Technische Universität Berlin

Berlin, Fachgebiet Datenbanksysteme und

Microsoft

Microsoft Deutschland GmbH

PTV GROUP

PTV GROUP - PTV Planung Transport Verkehr AG

SAP

SAP AG

SIEMENS

Siemens Aktiengesellschaft

software

SOFTWARE AG

StratoSphere

Stratosphere

TRUMPF

TRUMPF Werkzeugmaschinen GmbH

VOLKSWAGEN

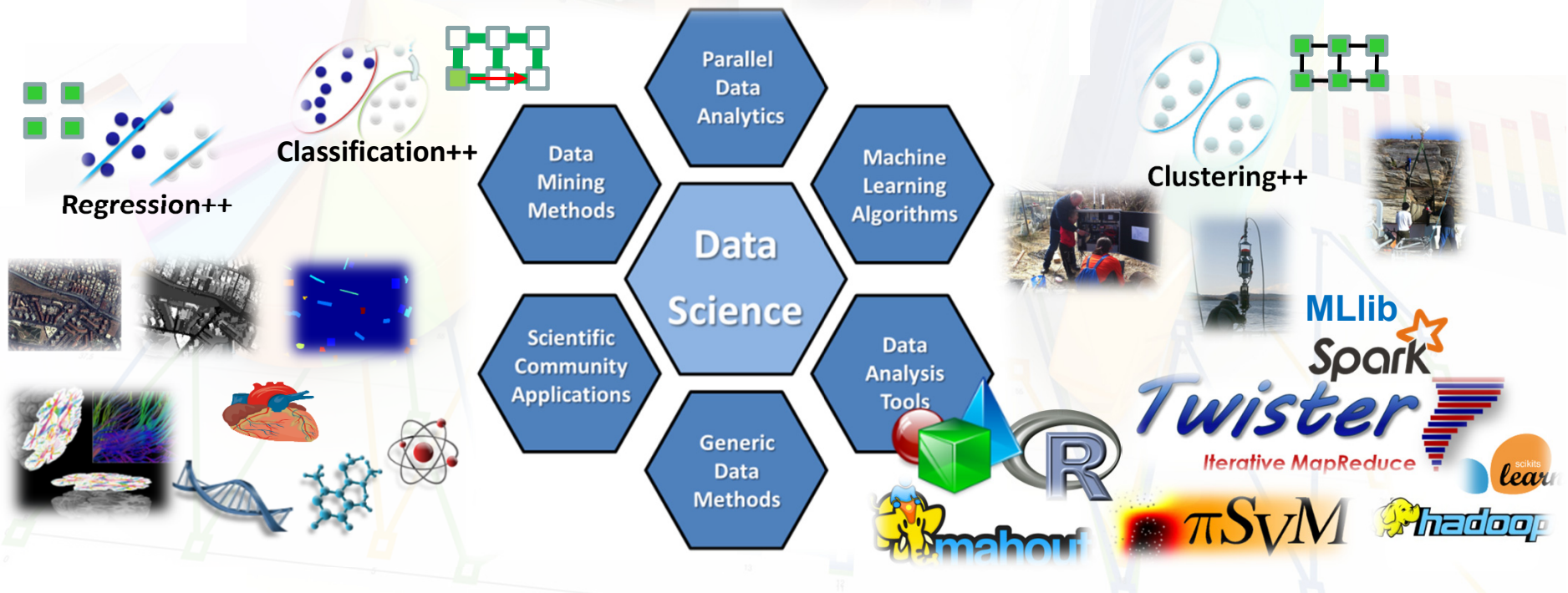
Volkswagen AG

Smart Data Innovation Lab



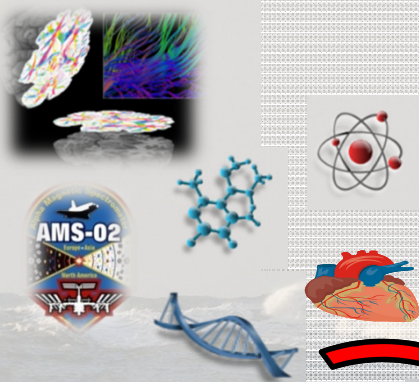
Serial Algorithms for Large Volumes of Data Exist

'Big Data' Requires Parallel Algorithms and Open Availability



Scientific Big Data Analytics: 'Big Data'-driven Research

Computation & Data Analysis gets more tightly intertwined

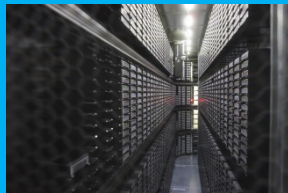


Scientific Computing



Scientific Applications using 'Big Data'

- Traditional Scientific Computing Methods
- HPC and HTC Paradigms & Parallelization
- Emerging Data Analytics Approaches
- Optimized Data Access & Management
- Statistical Data Mining & Machine Learning



'Big Data' Methods



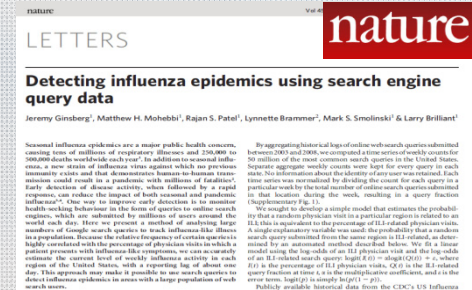
Scientific Applications



2009 – H1N1 Virus Made Headlines

Nature paper from Google employees
Explains how Google is able to predict winter flus
Not only on national scale, but down to regions
Possible via logged big data – ‘search queries’

[4] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, *Nature* 457, 2009

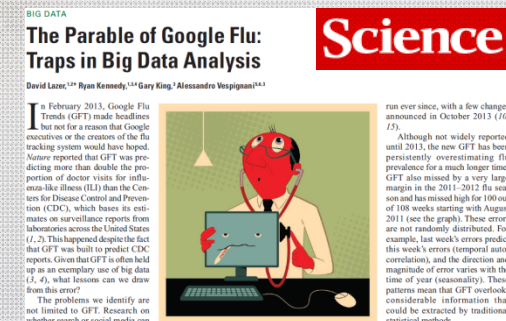


‘Big Data is not always better data’

2014 – The Parable of Google Flu

Large errors in flu prediction & lessons learned
(1) Dataset: Transparency & replicability impossible
(2) Study the algorithm since they keep changing
(3) It’s not just about size of the data

[5] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, *Science* Vol (343), 2014



Shifts from Causality to Correlation

Challenging research with progress based on reason?

'A clever combination of both is needed'



Traditional search for causality → (Big) Data Analysis

Exploring exactly WHY something is happening

Understanding causality is hard and time-consuming

Searching it often leads us down the wrong paths

Big Data Analytics

Not focussed on causality – enough THAT it is happening

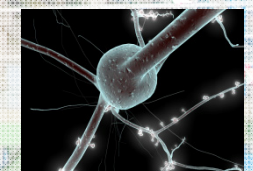
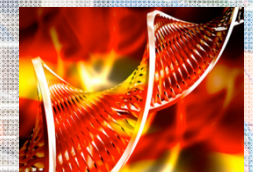
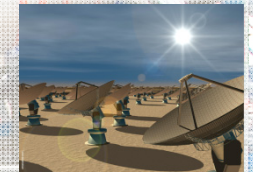
Discover novel patterns and WHAT is happening

Using correlations for invaluable insights – data speaks for itself



The Large Hadron Collider at CERN

'Scientific Big Data Application' with First Results



'Results today only possible due to extraordinary performance of Accelerators – Experiments – Grid computing'.

[7] Prof. Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery

Data Volume:

4 experiments / detectors

ca. 10^6 bytes / accident / experiment

ca. $3 \cdot 10^2$ accidents per second

ca. 10^5 seconds per day

ca. 10^2 (experiments-) days per year

→ $12 \cdot 10^{15}$ bytes / year

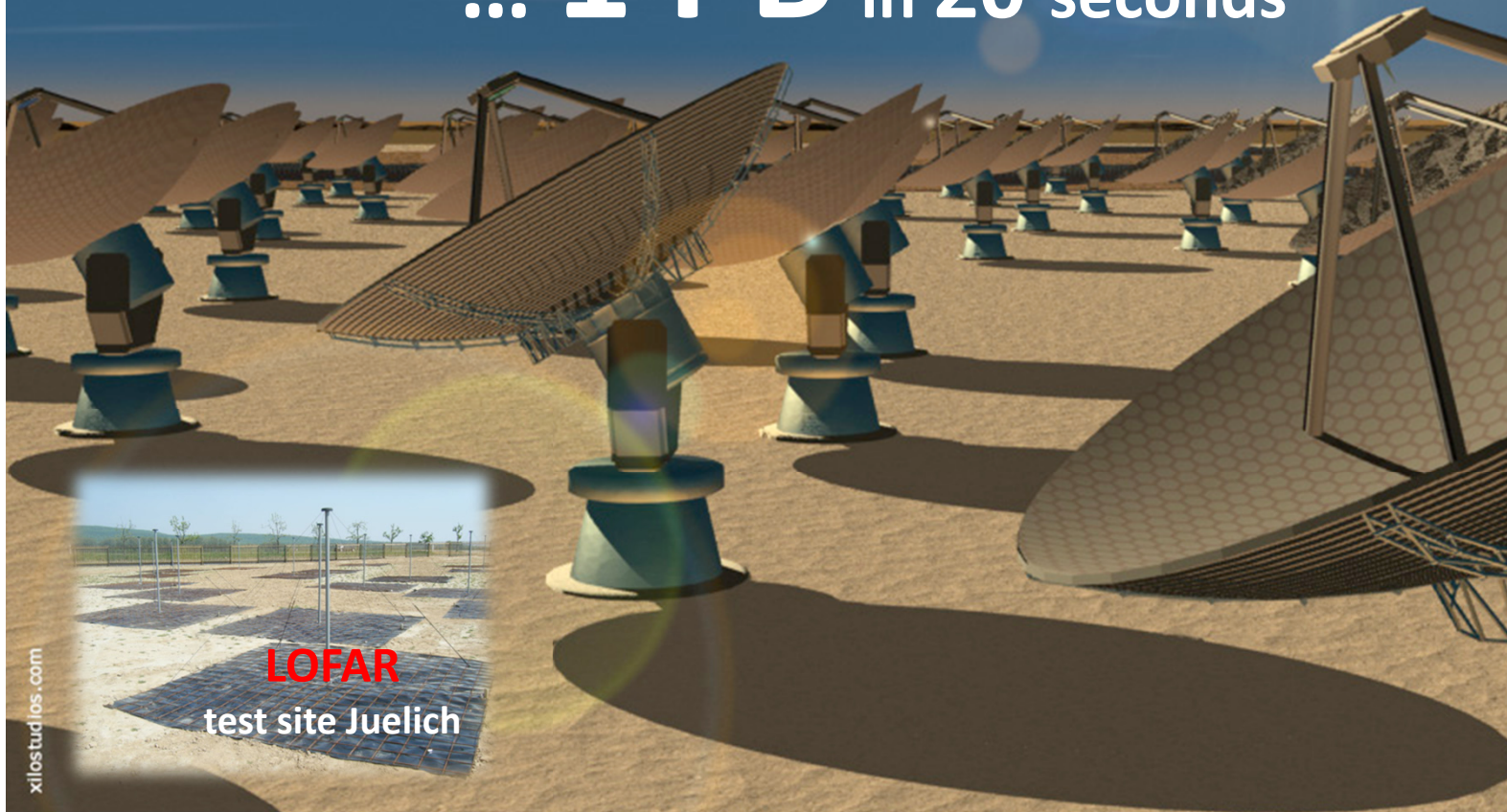
= 12 Petabytes per year

The next generation radio telescope for science...

... pushing the limits of the observable universe out by billions of galaxies

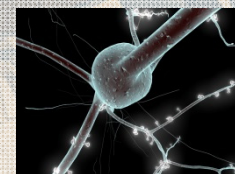
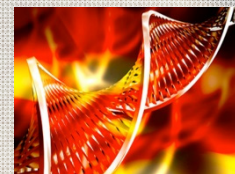
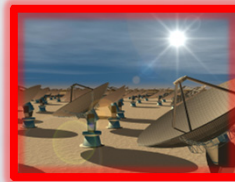
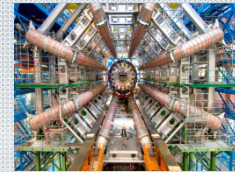
The Square Kilometre Array

... **1 PB** in 20 seconds

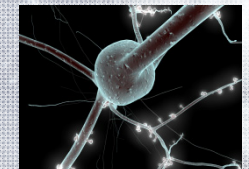
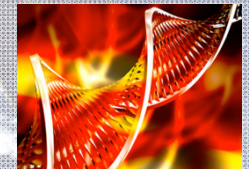
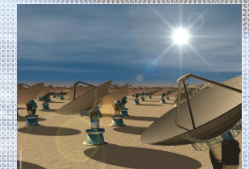
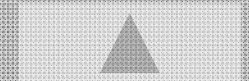


xilostudios.com

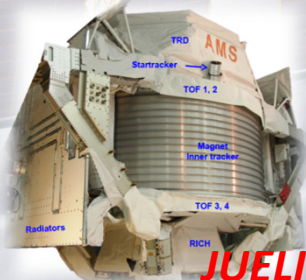
LOFAR
test site Juelich



What are building blocks of the Universe?



Alpha Magnetic Spectrometer (AMS) @ International Space Station



Search for
Cosmic
Antimatter

20
TB

~400
TB



JUROPA++

1-2 Pflop/s + Booster

JUELICH is main facility for AMS computing in Germany

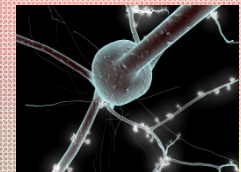
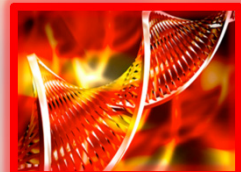
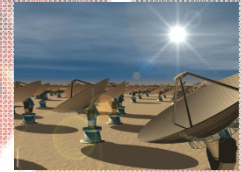
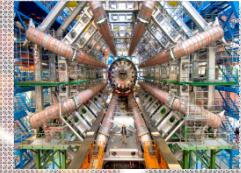
The 1000 Genome Project

Understanding what makes us different from one another



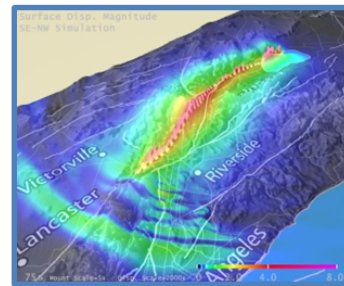
Comparing the complete DNA sequences of more than 1,000 individuals from around the world

**Next challenge: 2000 individuals
with each > 3 billion DNA base pairs
= 6 trillion DNA bases**



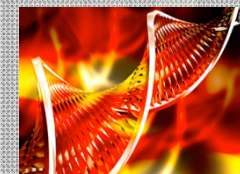
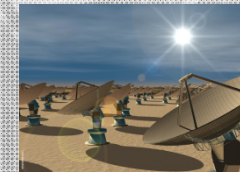
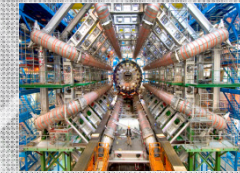
Large-scale Computational Parallel Applications Simulate Reality

<i>Estimated figures for simulated 240 second period, 100 hour run-time</i>	TeraShake domain (600x300x80 km ³)	PetaShake domain (800x400x100 km ³)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
Time steps	20,000 (.012 sec/step)	160,000 (.0015 sec/step)
Surface data	1.1 TB	1.2 PB
Volume data	43 TB	4.9 PB

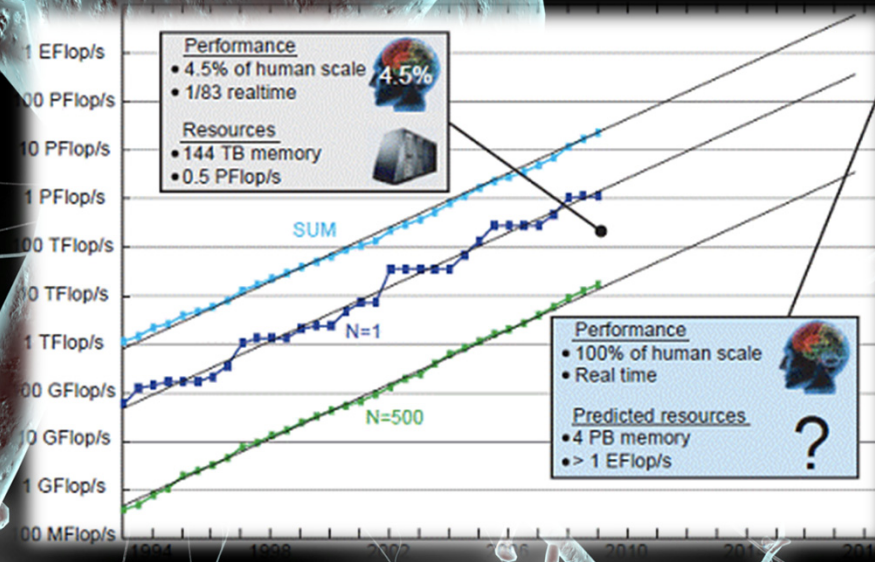


[8] Fran Berman, Maximising the Potential of Research Data

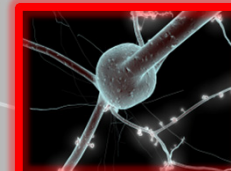
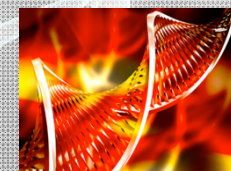
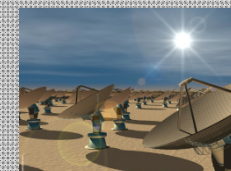
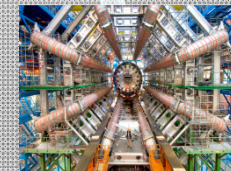
**Better Simulations...
... means 'Bigger Data'**



'A landing-on-the-moon-style project for neuroscience'



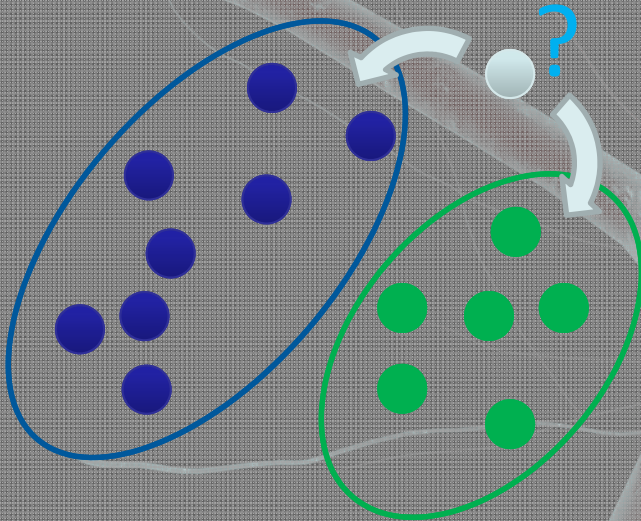
IP Human Brain Project



Complex Neuroscience Analytics

Lessons Learned towards Cloud and Autonomic Computing

Classification



Data Volume:

Block face images (of frozen tissue)

Every 20 micron (cut size)

Resolution: 3272 x 2469

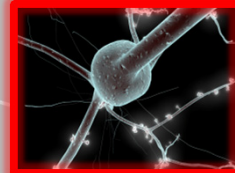
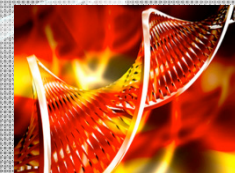
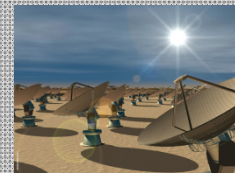
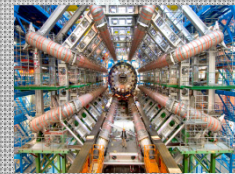
~14 MB / RGB image

~ 8 MB / corresponding mask image

~700 Images

→ **~40 GB dataset**

- Scientific Case: Understanding 'Sectioning of the brain'
- Goal: Build 'reconstructed brain (one 3d volume)' that matches with sections based on block face images



Model Selection and Cross-Validation

Working with Analytics Methods is a Process with Phases

Parallel Support Vector Machine

One of the most succesful classification methods

Classifier separates Two classes (brain, non-brain)

Parameters C & gamma after cross-validation

Cross-validation (grid-search) nicely parallel (e.g. for clouds)

Uses quadratic programming & Lagrangian method with **N x N**

$$\min_{w, \xi_i, b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \right\}$$

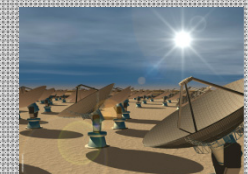
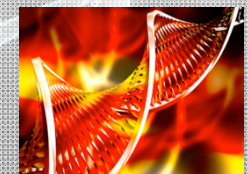
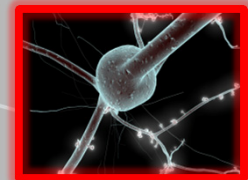
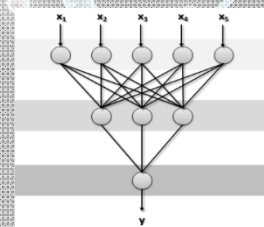
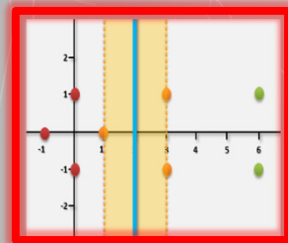
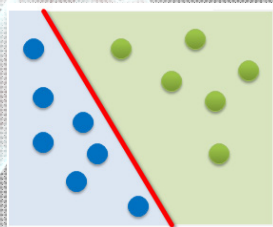
(optimization problem)

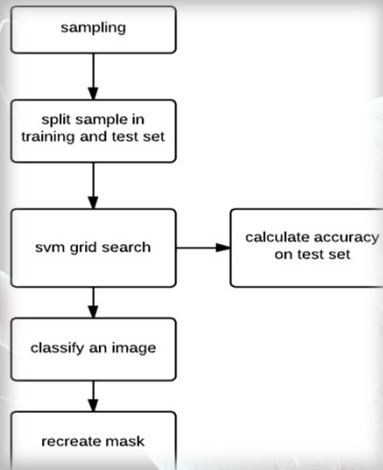
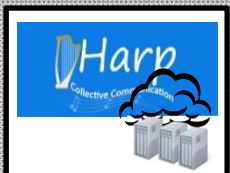
$$\mathcal{L}(\alpha) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N y_n y_m \alpha_n \alpha_m \mathbf{x}_n^T \mathbf{x}_m$$

(max. hyperplane → dual problem)

$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots & y_1 y_N x_1^T x_N \\ \dots & \dots & \dots & \dots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \dots & y_N y_N x_N^T x_N \end{bmatrix}$$

(quadratic coefficients)





Challenge:

Sampling bias:
more black than brain
pixels in groundtruth

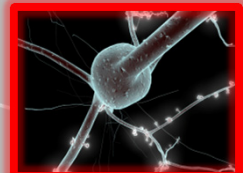
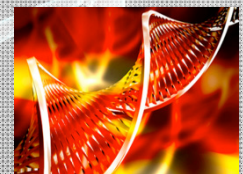
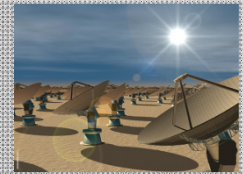
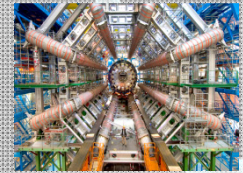
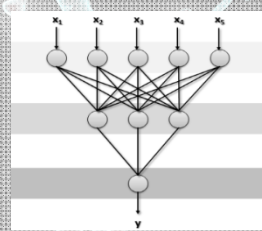
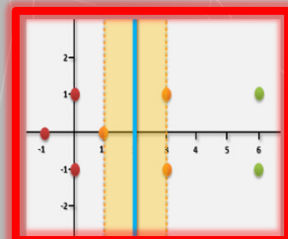
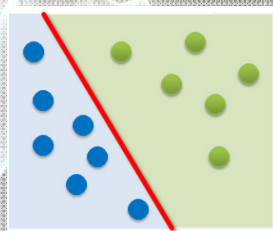


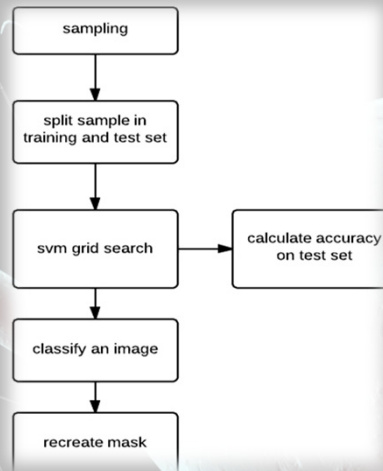
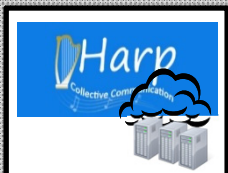
Sampling

Draw balanced random
class samples, eg.: 0.01%

Results

After performing manual sampling
Size of training data is limited (serial method)
→ 96% accuracy – but some remaining issues





Challenges:

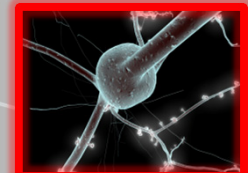
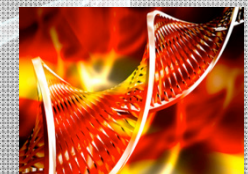
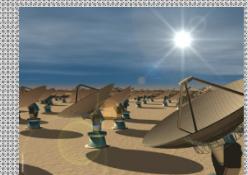
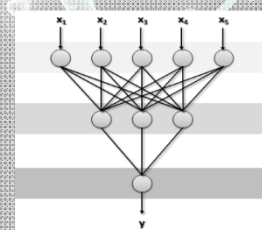
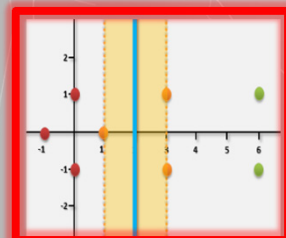
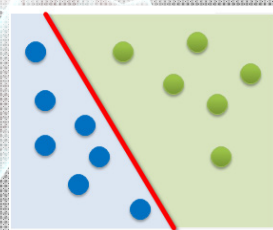
Distribute the data across the parallel infrastructure

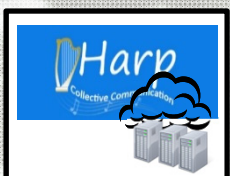
ParallelSVM implementation on top of Twister stack (~development version, but works)

Results

[15] Sun Z., and Fox G.,
'Study on Parallel SVM Based on MapReduce'

After performing manual sampling
Size of training data scales much better (parallel)
→ 96% accuracy – but some remaining issues



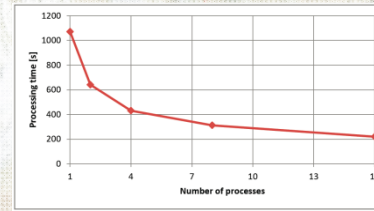
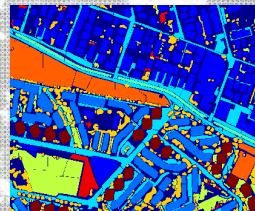


Remote Sensing Community

Large hyper- and multi-spectral datasets

Challenge: Multi-class classification

Classify different land-cover types



Results

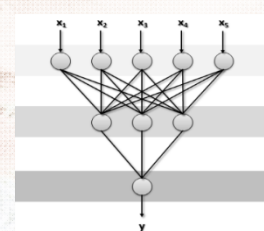
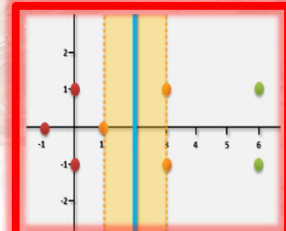
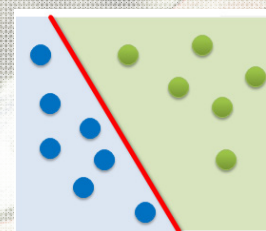
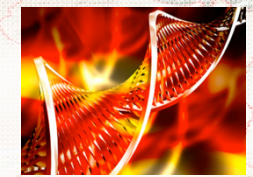
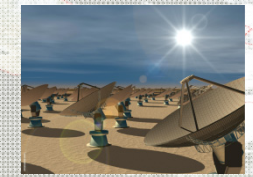
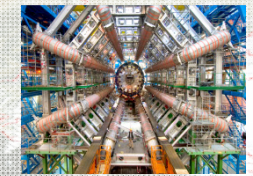
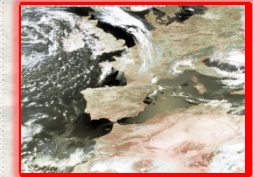
[12] G. Cavallaro and M. Riedel, 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014

Speed-up and decrease of training time

Applied Self-Dual Attribute Profile (SDAP)

From big data to smart data with statistics (PCA)

→97% accuracy



Big Data Applications – Statistical Data Mining Techniques

What is the right equipment, tool, technology, infrastructure?



The equipment and workbench used by Otto Hahn (1879 - 1968) and Fritz Strassmann in December 1938



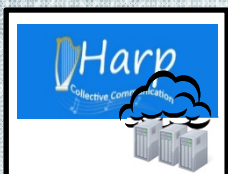
simple – yet powerful

'provided the first chemical evidence of nuclear fission products'

No opportunity to use an e-Infrastructure or Clouds

Big Data Technology is Available

But Need More Parallel Machine Learning

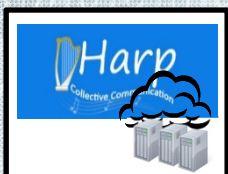


Our Workbench (e.g. focus on available parallel SVMs)

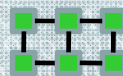
Tool	Platform Approach	Parallel Support Vector Machine
Apache Mahout	Java; Apache Hadoop 1.0 (map-reduce); HTC	No strategy for implementation (Website), serial SVM in code
Apache Spark/MLlib	Apache Spark; HTC	Only linear SVM; no multi-class implementation
Twister/ParallelSVM	Java; Apache Hadoop 1.0 (map-reduce); Twister (iterations), HTC	Much dependencies on other software: Hadoop, Messaging, etc.
Scikit-Learn	Python; HPC/HTC	Multi-class Implementations of SVM, but not fully parallelized
piSVM	C code; Message Passing Interface (MPI); HPC	Simple multi-class parallel SVM implementation outdated (~2011)
GPU accelerated LIBSVM	CUDA language	Multi-class parallel SVM, relatively hard to program, no std. (CUDA)
pSVM	C code; Message Passing Interface (MPI); HPC	Unstable beta, SVM implementation outdated (~2011)

Availability goes Beyond just 'Open Data'

Open Parallel Algorithm Implementations



Clustering++



Classification++



Regression++



Algorithm A Implementation

Algorithm Extension A' Implementation

Parallelization of Algorithm Extension A' \rightarrow A''

implementations available

implementations rare and/or not stable

closed/old source, also after asking paper authors



Towards Systematic Data Analytics

Guided by the Cross Industry Standard Process
for Data Mining (CRISP-DM) Phases

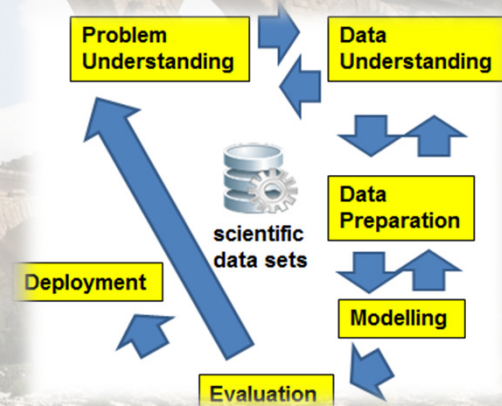
'Building a UCI Repository for Big Data Analytics'



RESEARCH DATA ALLIANCE

Big Data Analytics IG
Big Data Infrastructure WG

[10] Research Data Alliance



[11] P. Chapman et al., CRISP-DM Guide

„Reference Data Analytics“
for reusability & learning

CRISP-
DM
Report



Openly
Shared
Datasets



Running
Analytics
Code



Analytics Example



RESEARCH DATA ALLIANCE

Big Data Analytics IG
Big Data Infrastructure WG

[10] Research Data Alliance

Future
Grid

Twister
Iterative MapReduce

π SVM

Parallel
Brain Data
Analytics

scikits
learn



„Reference Data Analytics“
for reusability & learning

CRISP-
DM
Report



Openly
Shared
Datasets



Running
Analytics
Code



Sattelite Data(Quickbird)

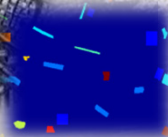
Parallel
Support Vector
Machines (SVM)

π SVM

HPC/MPI,
Map-Reduce &
GPGPUs



Classification
Study of
Land Cover
Types



‘Best Practices’

Community-
based practice

[12] G. Cavallaro and M. Riedel, ‘Smart Data Analytics
Methods for Remote Sensing Applications’, IGARSS 2014



[13] Open Grid Forum

**Basic Execution
Service**

**Job Submission
Description Language**

**GLUE2
Distributed Resource
Descriptions**

**Open
Cloud
Computing
Interface**

**Simple API for
Grid Applications**

**Distributed Resource
Management
Application API**

❖ *OGF is co-located with this
conference – check the program*



**Reliable Specifications to Build Upon
Standardized Building Blocks – ‘Rock Solid’**

Earth Science Data Analytics Examples

Take Advantage of Interoperability...

...between PRACE & XSEDE



[13] Open Grid Forum



[14] UNICORE.eu



PANGAEA

Problem: Quality control via outlier detection with PANGAEA data Collection



IAGOS

Problem: Longitude, latitude, altitude correlations with IAGOS data collection



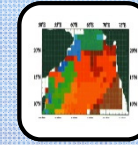
SCALE GIS

Problem: Projecting & transforming geospatial big data into a common coordinate reference framework



SEISMIC

Problem: Continuous seismic waveforms analysis for earthquakes monitoring



EVENTS

Problem: Event tracking analytics with spatial computing datasets (changing geolocations)

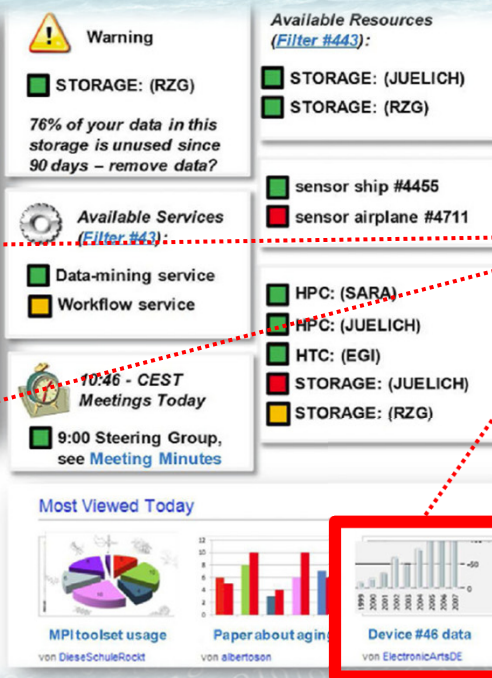


Understanding Possible Revenue Streams in Science & Engineering

Big Data Based Market-places

Enabling 'apps', 'subscription fees', 'advertisement', 'pay per use services'

- ❖ Hooks for offerings around commercial software packages
- ❖ Products around visualization packages and dedicated viewers
- ❖ Easy links to 'added value data', e.g. available market statistics
- ❖ Hosting services or deliver expandable storage in 'peek'
- ❖ Seamless links to the publishing industry



Data (or ScienceTube) to 'dive into data' with the possibility of commercial 'hooks'



[6] M. Riedel and P. Wittenburg et al. 'A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service', 2013

'In-Situ Analytics'





Methods & Tools

**Sampling
vs. Big Data**

Parallelization!

Applied Statistics

Data Mining

**Machine
Learning
Algorithms**

Scientific Computing



new DBs

Training Data Scientists

**Computational
Scientist**

**Software
Engineer**

Engineer

**Data
Miner**

Statistician

Data Scientist

Insights



UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING,
MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Statistical Data Mining Course

HPC – B(ig Data) Course

HPC – A(advanced) Scientific Computing Course

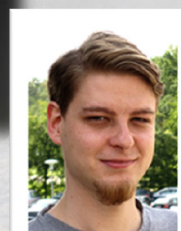
Data Scientists with skills of various fields



Acknowledgements

Selected Members of the Research Group on High Productivity Data Processing

Ahmed Shiraz Memon
Mohammad Shahbaz Memon
Markus Goetz
Christian Bodenstein
Philipp Glock
Matthias Richerzhagen



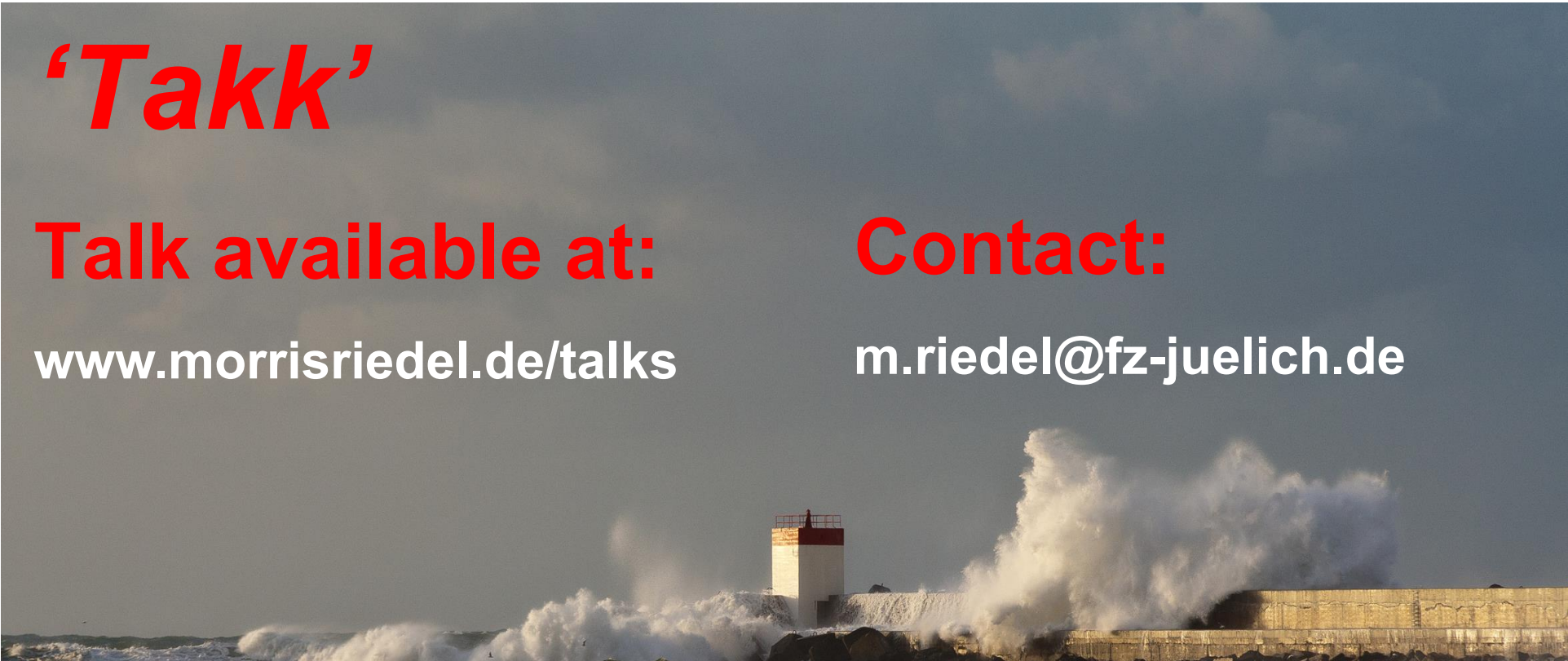
‘Takk’

Talk available at:

www.morrisriedel.de/talks

Contact:

m.riedel@fz-juelich.de

- 
- [1] Riding the Wave, EC Report, 2010
 - [2] A Surfboard for riding the wave, Report, 2012
 - [3] DoE ASCAC Report, 2013
 - [4] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, Nature 457, 2009
 - [5] David Lazer, Ryan Kennedy, Gary King & Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, Science Vol (343), 2014
 - [6] M. Riedel and P. Wittenburg et al. ‘A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service’, 2013
 - [7] Prof. Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery
 - [8] Fran Berman, Maximising the Potential of Research Data
 - [9] Study on Hadoop, Accenture
 - [10] Research Data Alliance, Big Data Analytics IG, Online: <https://rd-alliance.org/internal-groups/big-data-analytics-ig.html>
 - [11] P. Chapman et al., CRISP-DM Guide
 - [12] G. Cavallaro and M. Riedel, ‘Smart Data Analytics Methods for Remote Sensing Applications’, 35th Canadian Symposium on Remote Sensing (IGARSS), 2014, Quebec, Canada
 - [13] Open Grid Forum, Online: <http://www.ogf.org>
 - [14] UNICORE Technology, Online: <http://www.unicore.eu>
 - [15] Sun Z., and Fox G., ‘Study on Parallel SVM Based on MapReduce’, In Proceedings of the international conference on parallel and distributed processing techniques and applications, 2012.
 - [16] Apache Mahout Tutorial, YouTube Video